Radiomics of computed tomography and magnetic resonance imaging in renal cell carcinoma – a systematic review and meta-analysis

Stephan Ursprung MD[1,2], Lucian Beer MD PhD[1,2,3], Annemarie Bruining MD[4], Ramona Woitek MD[1,2,3], Grant D Stewart MD PhD[2,5], Ferdia A Gallagher MD PhD[1,2], Evis Sala MD PhD[1,2]

Affiliations

[1] Department of Radiology, School of Clinical Medicine, University of Cambridge, Cambridge UK
[2] Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge UK
[3] Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Vienna, Austria
[4] Department of Radiology, Netherlands Cancer Institute, Amsterdam, Netherlands
[5] Department of Surgery, School of Clinical Medicine, University of Cambridge, Cambridge UK

Institutional Address

Department of Radiology
University of Cambridge School of Clinical Medicine
Box 218, Cambridge Biomedical Campus
Cambridge
CB2 0QQ
United Kingdom

Corresponding Author

Professor Evis Sala
Department of Radiology
University of Cambridge School of Clinical Medicine
Box 218, Cambridge Biomedical Campus
Cambridge
CB2 0QQ
United Kingdom

es220@cam.ac.uk

Tel: 0044 1223 746 440
Fax: 0044 1223 330 915

Article Type

Original Research

Keywords

Carcinoma, Renal Cell; Angiomyolipoma; Machine Learning; Quality Improvement; Systematic Review

Key Points

- Studies achieved an average Radiomics Quality Score of 10.8%. Common reasons for low Radiomics Quality Scores were unvalidated results, retrospective study design, absence of open science and insufficient control for multiple comparisons.
- A previous training phase allowed reaching almost perfect inter-rater agreement in the application of the Radiomics Quality Score.
- Meta-analysis of radiomics studies distinguishing angiomyolipoma without visible fat from renal cell carcinoma show moderate diagnostic odds ratios of 6.24 and moderate methodological diversity. Abbreviations

Abbreviations

AMLwvf: Angiomyolipoma without visible fat

ICC: Interrater Correlation Coefficient

ML: Machine Learning

RCC: Renal Cell Carcinoma

RQS: Radiomics Quality Score

QUADAS: Quality Assessment of Diagnostic Accuracy Studies

Abstract

Objectives: (1) To assess the methodological quality of radiomics studies investigating histological subtypes, therapy response and survival in patients with renal cell carcinoma (RCC) and (2) to determine the risk of bias in these radiomics studies.

Methods: In this systematic review, literature published since 2000 on radiomics in RCC was included and assessed for methodological quality using the Radiomics Quality Score. The risk of bias was assessed using the Quality Assessment of Diagnostic Accuracy Studies tool and a meta-analysis of radiomics studies focusing on differentiating between angiomyolipoma without visible fat and RCC was performed.

Results: Fifty-seven studies investigating the use of radiomics in renal cancer were identified, including 4590 patients in total. The average Radiomics Quality Score was 3.41 (9.4% of total) with good inter-rater agreement (ICC:0.96, 95%-CI:0.93-0.98). Three studies validated results with an independent dataset, one used a publically available validation dataset. None of the studies shared the code, images or regions-of-interest. The meta-analysis showed moderate heterogeneity among the included studies and an odds ratio of 6.24 (95%-CI:4.27–9.12; $p<0.001$) for the differentiation of angiomyolipoma without visible fat from RCC.

Conclusions: Radiomics algorithms show promise for answering clinical questions where subjective interpretation is challenging or not established. However, the generalizability of findings to prospective cohorts needs to be demonstrated in future trials for progression towards clinical translation. Improved sharing of methods including code and images could facilitate independent validation of radiomics signatures.

Introduction

Radiological practice relies largely on the subjective interpretation of imaging data by an expert radiologist. Reports will therefore be dependent on reader experience. Quantitative, reader independent imaging markers may supplement expert opinion and increase diagnostic, predictive and prognostic accuracy [1]. Radiomics includes a number of strategies aimed at converting medical images to quantitative, minable, high-dimensional data. These include histogram, texture and shape analysis, that extract information from imaging data which may not be visible to the human eye [2, 3]. In recent years, increased interest in the use of radiomics in oncological imaging has led to its application as a tool to derive diagnostic, predictive and prognostic information from routine clinical imaging [4]. Despite extensive use in research and reports linking CT and MR texture to lesion characterization, survival and perioperative outcome in a number of malignancies, translation into clinical practice has not yet occurred [5].

Renal cell carcinoma (RCC) is newly diagnosed in 338,000 patients annually worldwide and incidence varies widely with the highest incidence in Northern America, Europe, Australia and New Zealand [6]. Most countries have seen a rise in incidence over the past decades, which has been attributed to the increasing use of cross-sectional imaging and subsequent incidental diagnosis [7]. Increasing diagnosis of small renal masses carries the risk of overtreatment resulting in benign histology in 10% - 30% of all resected tumors [8, 9]. While CT is the mainstay of diagnostic imaging in RCC, MRI has become a valuable problem-solving tool. Owing to its improved soft-tissue-contrast, MRI outperforms CT in the evaluation of indeterminate cystic masses (Bosniak 2F and 3, malignancy in 10% and 50% respectively) [10], local invasion and intra-vascular extension [11]. Still, the differentiation of benign renal lesions, especially oncocytoma and angiomyolipoma without visible fat (AMLwvf), from RCC can be challenging by subjective radiological image interpretation [12]. Quantitative image analysis may reveal radiomic signatures diagnostic of renal tumor subtype and aggressiveness or predictive of response to targeted treatment, therefore, aiding treatment stratification. However, for imaging markers including texture-based metrics to cross the translational gap between an exploratory research tool and a clinically applicable diagnostic algorithm, technical validity, biological validity, qualification, and cost-effectiveness need to be established (Figure 1) [13].

This systematic review aims to establish whether the methodological quality of prospective and retrospective studies published on radiomics in cross-sectional imaging of renal tumors for diagnostic, predictive and prognostic purposes poses barriers to effective clinical translation. A meta-analysis of the use of texture-based models for the discrimination of AMLwvf and RCC shall assess the ability of proposed models to answer this clinically relevant question.

Methods

This systematic review was conducted according to the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis for Diagnostic Test Accuracy) statement [14]. The review protocol is available through PROSPERO (CRD 42018115263). The electronic databases PubMed, EMBASE and Web of Science were searched for primary publications in English assessing texture analysis in RCC in CT or MRI published after 01/01/2000. The databases were last searched on the 30/10/2018. The search term consisted of (textural OR radiomics OR texture OR histogram) AND (kidney OR renal) AND (“computed tomography” OR CT or “magnetic resonance” OR MRI OR MR).

A single researcher with two years of post-graduate experience in medical image analysis (SU) screened titles and abstracts to determine eligibility. Articles in which texture analysis was employed for diagnostic, predictive or prognostic purposes on CT or MR images of RCC were obtained in full for further

evaluation. Contact with the authors was sought if the full-text version was not accessible otherwise. Studies were excluded if they were case reports, conference abstracts or short communications because they do not provide sufficient information to assess the methodological quality. The reference lists of included studies were screened for additional, potentially eligible articles. Uncertainties were resolved in consensus between SU, LB and AB

The Radiomics Quality Score (RQS) and the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) were used to assess the methodological quality of the included studies and the risk of bias on the study level, respectively [15, 16]. The RQS is a recently proposed tool to measure the methodological rigor of radiomics studies. It interrogates image acquisition, radiomics features extraction, data modelling, model validation, and data sharing. Each of the 16 dimensions (Table 1) of the score is rated resulting in a sum of points ranging from -8 to 36 with -8 – 0 defined as 0% and 36 defined as 100% [15]. The QUADAS-2 tool assesses the presence of bias in the domains of “patient selection”, “index test”, “reference standard” and “flow and timing”. The tool can be tailored to the specific research question through signaling questions for risks of bias which are specific to the individual research question [16].

During a training phase, the three reviewers (doctoral student with 2 years of postgraduate experience in medical image analysis (SU), a radiologist in the 4th year of training (LB) and a board-certified radiologist with 8 years of experience (AB)) independently extracted study data from two randomly chosen articles into a structured data collection instrument generated based on RQS and QUADAS-2. Disagreements were discussed in order to achieve a shared understanding of each parameter. Subsequently, at least two raters assessed and rated each study independently and recorded these on the data collection instrument. The data collection instrument can be found in supplementary Table S1.

Statistical analysis was conducted using R language for statistical computing [17]. Analyses were performed using the metafor, irr and raters packages [18]. Unless otherwise specified, the average rating of all raters is reported. Interrater agreement for single items of the RQS was calculated using a modified Fleiss kappa statistic for ordinal variables [19]. A 95% confidence interval was derived from a Monte Carlo test and bootstrap procedure over 1000 iterations. P-values for the null-hypothesis that agreement resulted from chance alone were calculated. The interclass correlation coefficient (ICC) was determined to describe interrater agreement for the summed RQS using a single source, two-way random effects model determining absolute agreement between raters.

As pre-defined in the review protocol, studies would be included in a meta-analysis of a large enough subset of the included studies if a similar clinical question was assessed repeatedly. Upon review of the study population, the differentiation of lesions defined as either fat poor AML, AMLwvf or AML without

macroscopic fat from malignant renal tumors was addressed repeatedly. These studies were included in the meta-analysis. Two-by-two contingency tables were extracted or reconstructed and odds ratios were calculated as effect size. A random-effects model was used to calculate the summary effect size. If multiple texture models were reported in a study, only the one with the highest area under the receiver operating curve or the highest Youden's J statistic, if no AUC was reported, was included. If data augmentation, the generation of new data through random transformation of existing cases, was performed, the augmented cases were not included in the meta-analysis. A funnel plot was constructed to visually assess the risk of publication bias and the trim and fill method was used to estimate the number of missing studies. Q and I^2 were calculated to estimate the heterogeneity among the studies included in the meta-analysis. A more detailed description of the statistical methods can be found in the supplementary materials.

Results

The initial search yielded 776 articles of which 263 were duplicates. Of the remaining 513, 454 were rejected based on title and abstract. Of the 59 full-text manuscripts retrieved, 57 were included in the systematic review (Figure 2). The articles employed radiomics-based diagnostic models to assess similar clinical questions repeatedly. The differentiation of benign and malignant lesions was investigated by 39% (22/57) of the articles while 27% (15/57) explored subtype differentiation and 21% (12/57) interrogated treatment response/outcome prediction. Tables S2 and S3 summarize study aims and characteristics, respectively.

The 57 studies reached a mean ± standard deviation RQS of 3.41±4.43, median 4.5, interquartile range: 6.17, range: -4.0-16.6. The average percentage RQS was 9.4% with a maximum of 46%. The average rating for each dimension is summarized in Table 1, the RQS for individual studies and individual ratings for each study are presented in Tables S2 and S4 respectively. Most studies applied discrimination statistics, included biological correlates and addressed their potential clinical utility. Conversely, none of the studies included in this systematic review employed phantoms or assessed the cost effectiveness of radiomics-based decision support systems. No study shared either segmentations or code publicly and only few assessed the repeatability of radiomics analysis at multiple time points, employed calibration statistics or a validation cohort. Only 39% (22/57) of the studies segmented the entire 3D tumor volume for texture analysis, and 91% (52/57) used manual segmentation. Inter-reader agreement was assessed in 32% (18/57) of the studies and found to be moderate to excellent for single features or radiomics signatures. Only a single study investigated the repeatability of radiomics measurements and found poor to good repeatability of histogram parameters of the transfer constant of dynamic contrast-enhanced MRI [20].

Studies included in this review extracted between four and 18,720 features (median 24) from two to 249 patients (median 61). The ratio between features and patients ranged from 25 times more patients than features to 240 times more features than patients (median of 2.2-times more patients than features). Feature reduction or adjustment for multiple testing was used in 51% of studies (29/57) and while 14% (8/57) relied on prospectively acquired data, none included plans for radiomics analysis in its prospective registration. Validation of radiomics signatures on independent validation datasets was performed in 5% (3/57) of the studies, only one of which employed an external dataset.

Assessment of the studies with the QUADAS-2 tool revealed methodological aspects increasing the risk of bias. As QUADAS-2 is not intended as a quantitative score, concern of bias from the reviewers was aggregated qualitatively for the different dimensions addressed by the tool (Table S5). Risk factors for bias which were repeatedly identified are summarized in Figure 3. Risk factors relating to patient selection and timing of index and reference tests were particularly frequently observed. Reporting the temporal delay between the index and reference test may be critical when determining tumor nuclear grade which influences progression and less critical when comparing RCC histological subtypes. The heavy reliance of literature on radiomics in RCC on retrospective surgical cohorts scanned with multiple scanners risks sampling technically variable data. Most studies explained texture feature extraction in detail; however, machine learning based models were employed in many papers without sufficient description of the model parameters to allow replication.

The reproducibility of the RQS and QUADAS-2 was also assessed. During the training phase, particular variability in the rating of the detection and discussion of biological correlates was identified. The reviewers agreed to rate the item more liberally in agreement with previous publications [21]. The ICC for the RQS was 0.96 (95%-CI:0.93-0.98). The ICC for studies rated by all three reviewers (11/57) was 0.92 (95%-CI:0.80-0.98). Substantial or almost perfect agreement was achieved for most individual elements of the RQS. Only moderate agreement was reached in the assessment of the imaging protocol (Table 2). Absolute agreement concerning risk of bias and applicability of the seven indicator questions of the QUADAS tool was generally above 75% for most dimensions. Absolute agreement was 58% in the assessment of the risk of selection bias.

Publication bias is a concern in radiomics studies in particular. Indeed, only 4/57 (7%) publications included in this review report non-significant outcomes, all analyzing the differentiation of AML and RCC. In the absence of prospective investigations with predefined study protocols, selective reporting of positive outcomes is a risk.

Thirteen of the 57 studies (23%) discussed the use of radiomics for the differentiation of AMLwvf and malignant renal tumors. Of these, 77% (10/13) provided information to reconstruct a contingency table and calculate the effect size and were included in the meta-analysis. The summary effect size under the random effects model across the studies indicated a diagnostic odds ratio of 5.89 (95% CI: 4.02-8.23 p<.001) for radiomics models differentiating AMLwvf from RCC (Figure 4). Cochran's Q of 13.41, p=.15 with 9 degrees of freedom and I^2=33.5% suggested the presence of moderate study-to-study dispersion. The funnel plot relating effect size to its standard error is shown in Figure 5. Trim and fill analysis estimated that one study on the left side was missing. Following the addition of this study, the estimated overall effect size is OR=5.55 (95% CI: 3.77-8.16, p<0.001). Considerable diversity existed among the radiomics features calculated and only mean in the unenhanced, entropy in the unenhanced and nephrographic phase CT were found to differentiate AMLwvf and RCC in two studies. Two studies assessing the ability of low attenuation voxel percentage to differentiate AMLwvf and RCC found significant differences in opposing directions [22, 23].

Discussion

Radiomics may provide new quantitative imaging markers without the need to invest in new acquisition equipment or tracers. Multiple studies have shown promise in answering clinical questions that conventional, qualitative radiological diagnosis cannot answer. However, none of the multifactorial radiomics algorithms has achieved clinical translation or been independently validated. This systematic review has identified several common characteristics among the included studies that hinder rapid adoption of proposed algorithms into the clinic. Replication and independent validation of research findings relies on sharing of imaging data, segmentations and code. None of the studies included in this review have provided open access to the code employed for data preparation, feature extraction and model construction. This is particularly crucial where image pre-processing and artificial intelligence based modeling were applied. Guidelines recommending reporting standards for machine learning (ML) models have been published, however, making the code used for data analysis publically available would be preferable [24]. Overall, 34/57 studies used ML models. There was a trend for these studies to be more recent than those not using any ML models. Furthermore, studies incorporating ML algorithms achieved significantly higher RQS ratings than studies without (5.16 ± 3.66 vs. 0.83 ± 4.27, p < .001). This was due, in particular, to less frequent validation of results, inclusion of non-radiomics parameters and use of feature reduction and correction for multiple comparisons in non-ML studies.

Where patient numbers are limited and countless radiomics features can be quantified, it is critical to reduce the feature space, e.g. through removal of poorly reproducible features to reduce the risk of overfitting. This could be achieved with texture phantoms that were not employed by any of the studies

in this review. Furthermore, appropriate statistical correction for multiple comparisons and independent validation, which has only been applied very rarely among the included studies, will reduce the risk of false positive and overly optimistic results. Meanwhile, prospective trials, where hypotheses are defined in advance, reduce the risk of reporting bias. Most trials included in this review only assessed surgical patients. However, surgical cohorts may be enriched in malignant lesions and larger tumor sizes, leading to selection bias. Small renal masses, which can be difficult to classify, may be assigned to active surveillance and are, therefore, underrepresented in surgical cohorts.

The RQS has been proposed to assess the methodological quality of radiomics studies, which is important to critically appraise the large number of publications and to prioritize validation of high quality results. Because varying interrater agreement was observed in the first application of the RQS [21], two articles were used to train researchers. As a result, high agreement for the overall rating (ICC = 0.96) and most elements of the score (S* > 0.75) was achieved. Compared to the first application of the RQS, the average RQS rating was lower (10.8% vs 21.9%) as was the rating for the best performing study (48% vs 55.5%). Another, recently published review employing the RQS did not report interrater agreement [25].Only few systematic reviews in radiomics literature have been published and even fewer assessed methodological quality systematically and quantitatively As a result, the RQS has not yet found widespread application.

The dependency of multiple radiomics features on image acquisition parameters has been demonstrated repeatedly [26–28]. However, only half of the studies included in this review, documented the most important parameters. The selection of the scanner manufacturer and model, acquisition and reconstruction parameters cause heterogeneity of imaging data. If the aim is to achieve broadly applicable radiomics models, standardization will be required wherever possible. Elsewhere, feature selection could consider robustness to variations in acquisition parameters and adjustments could be applied to the input data or the extracted features. The non-quantitative nature of T_1 and T_2-weighted MR sequences introduces additional heterogeneity even when acquisition parameters are kept constant. As a result, MR-based radiomics models frequently employed parametric maps, which do not require initial signal intensity normalization, were used most commonly. Out of 17 studies using MR, nine included the advanced diffusion coefficient based on diffusion-weighted imaging and two the transfer constant k^{Trans} from dynamic contrast-enhanced MRI in their analysis. Two studies employed ADC histogram parameters to differentiate tumor subtypes observing similar trends but differential statistical significance due to low numbers of cases.

Most studies segmented only part of the tumor. In light of recent findings highlighting the intratumoral heterogeneity in RCC on a genetic and metabolic level, texture analysis in a single 2D slice risks underestimating intratumoral heterogeneity [29–31]. However, studies segmenting single 2D slices of the tumor achieved equal RQS ratings and no trend over time favoring one segmentation strategy was apparent. The few publications comparing 2D and 3D texture analysis reached varying conclusions regarding their ability to correctly measure heterogeneity in tumors. However, it seems premature to suggest that segmentation of single slices was equivalent in diagnostic value to segmentation of an entire lesion. Only a small subset of the studies (5/57) placed small regions of interest within the tumor. These were either very early studies or studies where multiregional tissue sampling to match the regions of interest was carried out. Additionally, there is scope for further integration of radiomics data with clinical, genetic and metabolic data to achieve a more complete understanding of renal cancer and harness the complementary value of each modality in cancer diagnostics, prognosis, treatment response prediction and monitoring.

This review has some inherent limitations. First, the articles included in the meta-analysis differed slightly in their inclusion criteria. The control group was composed of ccRCC only in four studies while six included RCC of multiple subtypes. The methodology will always differ between radiomics studies as different centers use different equipment and the choice of image reconstruction, filtration, feature extraction and calculation of radiomics models offer countless combinations. Still, a meta-analysis of the existing evidence provides important information as to the consistency of results and the magnitude of the effect size that can be anticipated and helps to estimate publication bias. Notably, the clinically more relevant question of differentiating oncocytoma from RCC was less frequently assessed. A number of studies included in this review were published before the introduction of the RQS. However, there was no trend for improvement over time, therefore this was not thought to be a significant risk of bias. The RQS as well as QUADAS-2 have limitations. While the former is a quantitative metric and a debate about the appropriate weighting of different components is justified, the latter is a qualitative score and therefore less easily interpretable. Still, both scores are timely tools for the assessment of methodological quality of this highly specialized area of research.

In conclusion, radiomics models show promise for augmenting radiological diagnosis in renal cancer. The differentiation of AMLwvf and RCC has been investigated repeatedly and a meta-analysis showed moderate ability of radiomics to facilitate this distinction. However, well-designed and appropriately powered prospective radiomics trials are needed for these novel imaging markers to demonstrate their validity and progress towards clinical translation.

References

1. Sullivan DC, Obuchowski NA, Kessler LG, et al (2015) Metrology Standards for Quantitative Imaging Biomarkers. Radiology 277:813–825
2. Castellano G, Bonilha L, Li LM, Cendes F (2004) Texture analysis of medical images. Clin Radiol 59:1061–9
3. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: Images Are More than Pictures, They Are Data. Radiology 278:563–577
4. Lubner MG, Smith AD, Sandrasegaran K, et al (2017) CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges. RadioGraphics 37:1483–1503
5. Miles KA (2016) How to use CT texture analysis for prognostication of non-small cell lung cancer. Cancer Imaging 16:10
6. Ferlay J, Soerjomataram I, Dikshit R, et al (2015) Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 136:E359–E386
7. Znaor A, Lortet-Tieulent J, Laversanne M, et al (2015) International Variations and Trends in Renal Cell Carcinoma Incidence and Mortality. Eur Urol 67:519–530
8. Pierorazio PM, Hyams ES, Mullins JK, Allaf ME (2012) Active surveillance for small renal masses. Rev Urol 14:13–9
9. Richard PO, Lavallée LT, Pouliot F, et al (2018) Is routine use of renal tumor biopsy associated with lower rates of benign histology following nephrectomy for small renal masses? J Urol 0: . doi: 10.1016/j.juro.2018.04.015
10. Defortescu G, Cornu J-N, Béjar S, et al (2017) Diagnostic performance of contrast-enhanced ultrasonography and magnetic resonance imaging for the assessment of complex renal cysts: A prospective study. Int J Urol 24:184–189
11. Karlo CA, Di Paolo PL, Donati OF, et al (2013) Renal Cell Carcinoma: Role of MR Imaging in the Assessment of Muscular Venous Branch Invasion. Radiology 267:454–459
12. Hindman N, Ngo L, Genega EM, et al (2012) Angiomyolipoma with Minimal Fat: Can It Be Differentiated from Clear Cell Renal Cell Carcinoma by Using Standard MR Techniques? Radiology 265:468–477
13. O'Connor JPB, Aboagye EO, Adams JE, et al (2017) Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol 14:169–186
14. McInnes MDF, Moher D, Thombs BD, et al (2018) Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies. JAMA 319:388
15. Lambin P, Leijenaar RTH, Deist TM, et al (2017) Radiomics: The bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14:749–762
16. Whiting PF, Rutjes AWS, Westwood ME, et al (2011) QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med 155:529
17. (2013) R: A language and environment for statistical computing - R foundation for Statistical Computing, Vienna, Austria. In: R Core Team. http://www.r-project.org/
18. Viechtbauer W (2010) Conducting Meta-Analyses in *R* with the metafor Package. J Stat Softw 36:1–48
19. Marasini D, Quatto P, Ripamonti E (2016) Assessing the inter-rater agreement for ordinal data

through weighted indexes. Stat Methods Med Res 25:2611–2633

20. Wang HY, Su ZH, Xu X, et al (2016) Dynamic Contrast-enhanced MR Imaging in Renal Cell Carcinoma: Reproducibility of Histogram Analysis on Pharmacokinetic Parameters. Sci Rep 6: . doi: 10.1038/srep29146

21. Sanduleanu S, Woodruff HC, de Jong EECC, et al (2018) Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. Radiother Oncol 127:349–360

22. Kim JY, Kim JK, Kim N, Cho K-S (2008) CT Histogram Analysis: Differentiation of Angiomyolipoma without Visible Fat from Renal Cell Carcinoma at CT Imaging. Radiology 246:472–479

23. Catalano OA, Samir AE, Sahani D V., Hahn PF (2008) Pixel Distribution Analysis: Can It be Used to Distinguish Clear Cell Carcinomas from Angiomyolipomas with Minimal Fat? Radiology 247:738–746

24. Jethanandani A, Lin TA, Volpe S, et al (2018) Exploring Applications of Radiomics in Magnetic Resonance Imaging of Head and Neck Cancer: A Systematic Review. Front Oncol 8:131

25. Park JE, Kim D, Kim HS, et al (2019) Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol 1–14

26. Shafiq-ul-Hassan M, Zhang GG, Latifi K, et al (2017) Intrinsic dependencies of CT radiomics features on voxel size and number of gray levels. Med Physiccs 44:1050–1062

27. Berenguer R, Pastor-Juan M del R, Canales-Vázquez J, et al (2018) Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. Radiology 288:172361

28. Zhao B, Tan Y, Tsai WY, et al (2014) Exploring Variability in CT Characterization of Tumors: A Preliminary Phantom Study. Transl Oncol 7:88–93

29. Gerlinger M, Rowan AJ, Horswell S, et al (2012) Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. N Engl J Med 366:883–892

30. Turajlic S, Xu H, Litchfield K, et al (2018) Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal. Cell 173:581–594

31. Okegawa T, Morimoto M, Nishizawa S, et al (2017) Intratumor Heterogeneity in Primary Kidney Cancer Revealed by Metabolic Profiling of Multiple Spatially Separated Samples within Tumors. EBioMedicine 19:31–38

Figure legends

Figure 1: Pathway for the development of radiomics algorithms and challenges in clinical translation. In addition to image acquisition and image registration, non-quantitative MRI sequences may undergo intensity normalization to improve intra- and inter-patient heterogeneity. Subsequently, either classical machine learning algorithms or deep learning are employed to define diagnostic, prognostic or predictive models. These models require external validation, ensuring transferability of results between sites and MR-scanners before prospective validation and demonstration of cost-effectiveness can enable these diagnostic support systems to enter clinical practice. Continuous monitoring is required to detect deteriorating model performance as image acquisition evolves to trigger re-training and model update. ANN: Artificial Neural Network

Figure 2: Study selection flowchart.

Figure 3: Risk factors for bias colored according the four dimensions of the QUADAS tool. The length of the bars is equivalent to the frequency with which this risk factor was identified among the included studies.

Figure 4: Forrest plot of the effect size calculated as log odds ratio for 10 of 13 studies investigating the diagnostic accuracy of radiomics in the differentiation of AMLwvf from RCC. TP: Number of AMLwvf patients correctly diagnosed, FN Number of AMLwvf patients diagnosed as having RCC. FP Number of RCC patients diagnosed as having AML, TN number of RCC patients correctly diagnosed. X-axis: log-transformed odds ratios, RE: random effects

Figure 5: Funnel plot of studies included in the meta-analysis (black) and missing studies identified by trim and fill analysis (white dot). The funnel plot was asymmetric with more studies than expected reporting higher odds ratios for the ability of radiomics to differentiate between malignant renal tumors and benign AMLwvf, this can indicate the presence of publication bias.

Table legends

Table 1 Elements of the RQS as described by Lambin *et.al.* [15] and average rating achieved by the studies included in this systematic review.

Table 2: Inter-Rater agreement in the assessment of the RQS

Figures

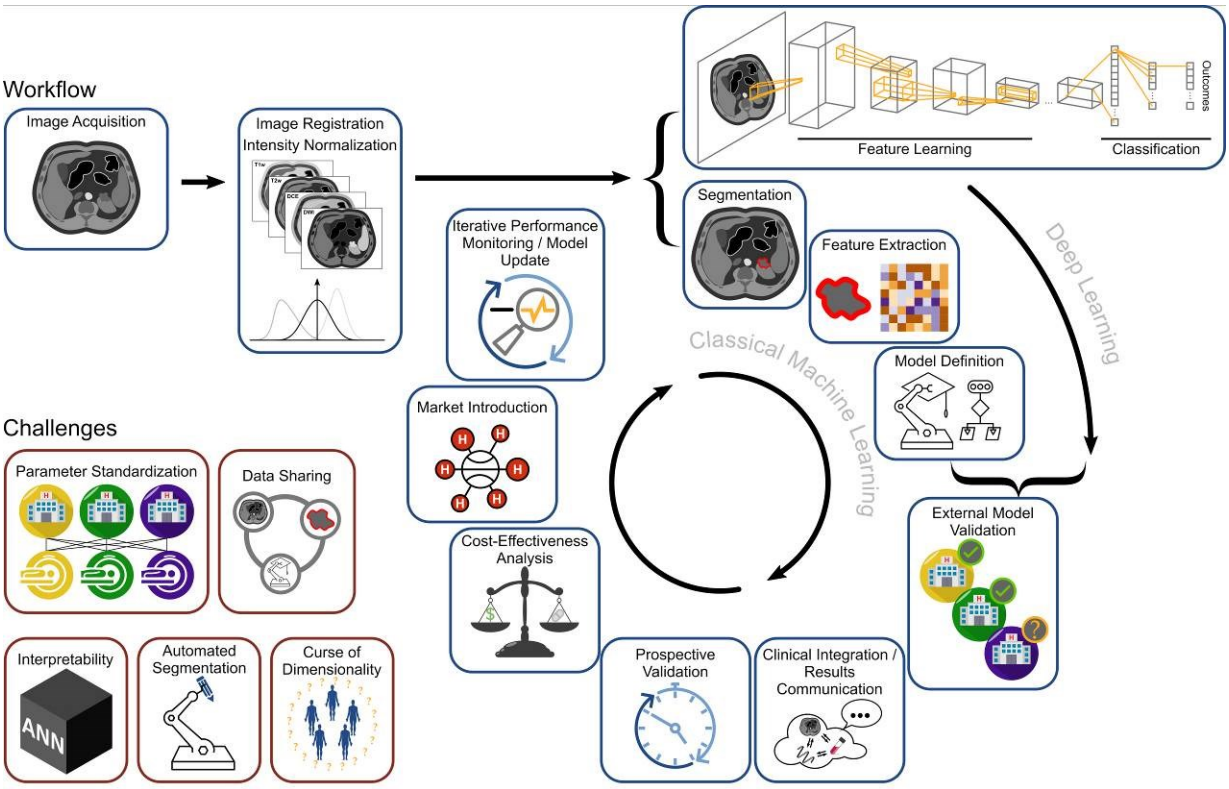

Figure 1

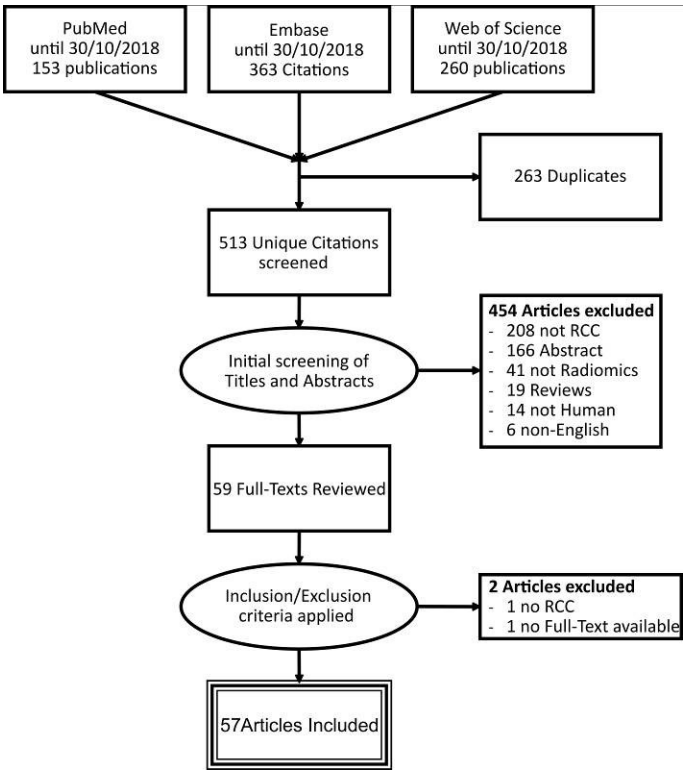

Figure 2

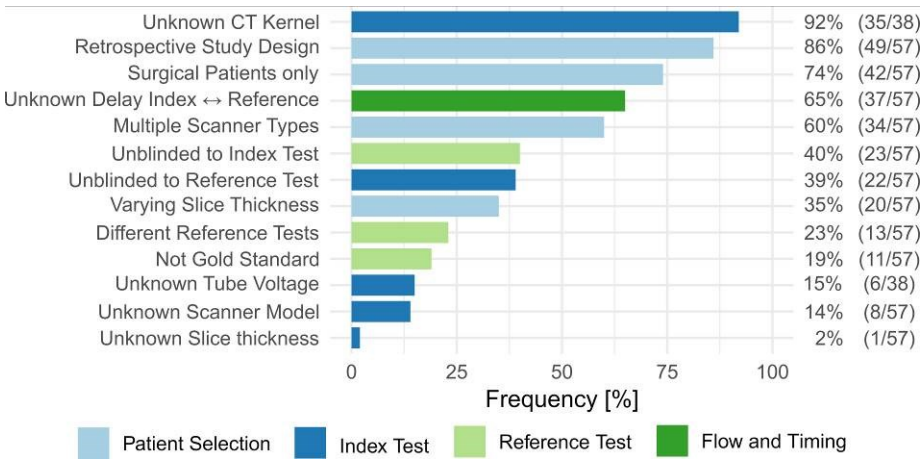

Figure 3

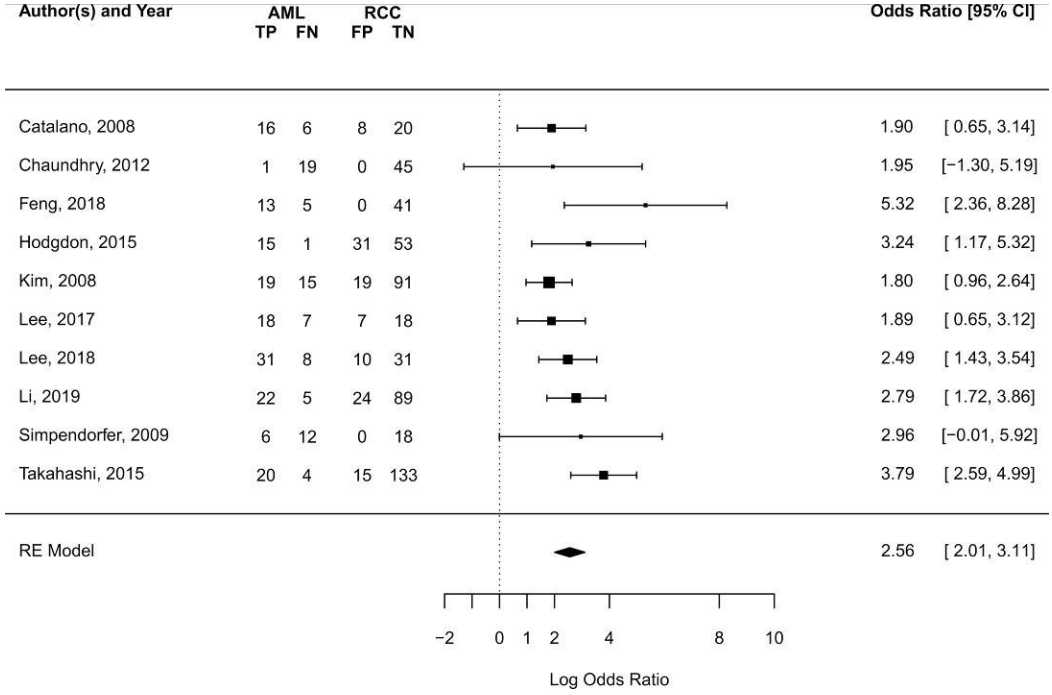

Figure 4

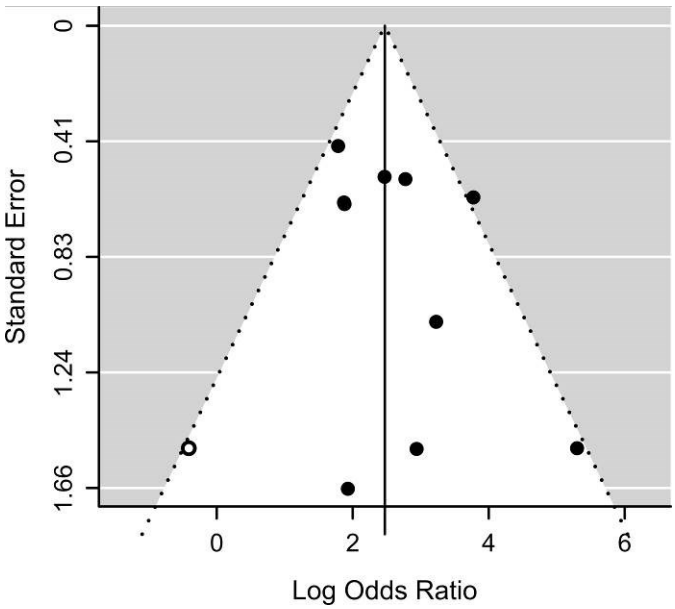

Figure 5

Tables

Table 1

RQS Scoring Item	Interpretation	Average
Image Protocol	+1 for well documented protocols, +1 for publicly available protocols	0.48
Multiple Segmentations	+1 if segmented multiple times (different physicians, algorithms or perturbation of regions of interest)	0.38
Phantom Study	+1 if texture phantoms were used for feature robustness assessment	0.00
Multiple Time Points	+1 multiple timepoints for feature robustness assessment	0.01
Feature Reduction	-3 if nothing, +3 if either feature reduction or correction for multiple testing	0.23
Non Radiomics	+1 if multivariable analysis with non-radiomics features	0.15
Biological Correlates	+1 if present	0.98
Cut-off	+1 if cutoff either pre-defined or at median or continuous risk variable reported	0.11
Discrimination and Resampling	+1 for discrimination statistic and statistical significance, +1 if resampling applied	0.92
Calibration	+1 for calibration statistic and statistical significance	0.04
Prospective	+7 for prospective validation within a registered study	0.98
Validation	-5 if no validation / +2 for internal validation / +3 for external validation / +4 two external validation datasets or validation of previously published signature / +5 validation on ≥3 datasets from >1 institute	-4.61
Gold Standard	+2 for comparison to gold standard	1.73
Clinical Utility	+2 for reporting potential clinical utility	1.91
Cost-effectiveness	+1 for cost-effectiveness analysis	0.00
Open Science	+1 for open source scans, +1 for open source segmentations, +1 for open source code, +1 open source representative segmentations and features	0.02

RQS: Radiomics Quality Score

Table 2

RQS Scoring Item	S* [95% CI]
Image Protocol	0.45 [0.20 – 0.67]
Multiple Segmentations	0.93 [0.82 – 1.00]
Phantom Study	1.00 [1.00 – 1.00]
Multiple Time Points	0.93 [0.82 – 1.00]
Feature Reduction	0.93 [0.82 – 1.00]
Non Radiomics	0.67 [0.49 – 0.85]
Biological Correlates	0.93 [0.82 – 1.00]
Cut-off	0.93 [0.82 – 1.00]
Discrimination and Resampling	0.82 [0.71 – 0.92]
Calibration	0.96 [0.89 – 1.00]
Prospective	1.00 [1.00 – 1.00]
Validation	1.00 [1.00 – 1.00]
Gold Standard	0.76 [0.63 – 0.88]
Clinical Utility	0.60 [0.38 – 0.82]
Cost-effectiveness	1.00 [1.00 – 1.00]
Open Science	1.00 [1.00 – 1.00]

CI: Confidence Interval, RQS: Radiomics Quality Score